Transposons are mobile genetic elements capable of random expression analysis (Judson and Mekalanos, 2000; Opperman et al, 1995), transposon-mediated differential hybridization (Chaudhuri et al, 2003; Yergeau and Mead, 2009). Various modifications of traditional transposon mutagenesis facilitated the development of new techniques, such as signature target mutagenesis (Hensel et al, 1999), transposon-mediated differential hybridization (Chaudhuri et al, 2009), genetic footprinting (Smith et al, 1999) and gene expression analysis (Hudson and Mekalanos, 2003; Opperman et al, 2001).

Here, we present a novel computational method for the identification of insertion sites in transposon-mutagenized bacterial genomes. The functionality of the algorithm was successfully demonstrated using a mini-Tn10 transposon Legionella pneumophila Corby library.

2 METHODS

2.1 InFiRe algorithm

To improve the efficiency of transposon mutagenesis experiments, we developed an algorithm for the identification of insertion sites in transposon-mutagenized bacterial genomes. The strategy of this procedure is outlined in Figure 1. It is based on the restriction digestions of genomic DNA in combination with Southern blot hybridization. In the first step, the genomic DNA is cleaved with different restriction endonucleases. Each restriction enzyme digest produces a unique pattern of DNA fragments with defined sizes. In the second step, the sizes of the fragments with an inserted transposon are determined by Southern blot hybridization with a transposon-specific probe. The size of the fragment in the original genome equals the obtained fragment size minus the length of the transposon. Finally, using the derived fragment size pattern the most probable genomic position of the transposon can be calculated by the InFiRe software. Afterwards, confirmation of the accurate determination of insertion sites can be performed by PCR.

2.2 Statistics

In order to determine the number of restriction enzymes required for InFiRe analyses, we developed two statistical models. The explicit model calculates the number of enzymes based on the genome sequence information, whereas the theoretical model estimates this number based on a random sequence of given length. Both models produce similar results.
InFiRe

Fig. 1. Procedure for the identification of the transposon insertion in bacterial genomes. Chromosomal DNA from a transposon insertion mutant (A) is digested with a set of restriction enzymes (B). Each restriction digestion results in a unique pattern of DNA fragments. Separation of the DNA fragments by agarose gel electrophoresis followed by Southern blot hybridization with a transposon-specific probe allows determination of the approximate size of chromosomal fragments containing the transposon (C). The pattern of the derived fragment sizes allows the calculation of the most probable genomic position of the transposon via the InFiRe software (D and E).

2.3 Explicit model

The outlined explicit statistics approach calculates the number of digestions and shows the most suitable combination of restriction enzymes with an error probability. In the first step, the fragment size distribution for each digestion \( i \) is calculated for a given genome sequence. Fragment sizes \( s \) are distributed almost exponentially. Assuming a random and unbiased DNA sequence, the approximated distribution \( P_1(s) \) of the fragment sizes is calculated by Equation (1) (Fig. 2).

\[
P_1(s > x) \sim e^{-x/m}, \quad \forall x \geq 0
\]

In the next step, the maximum achieved number of the fragments \( b_i \) at the size \( s_{\text{max}} \) is calculated for each digestion \( i \).

\[
b_i := \max \{ P_2(s) \} = P_2(s_{\text{max}}), \quad \forall x \geq 0
\]

Finally, for every \( i \) (ordered by \( b_i \)) all overlaps between the fragments in \( Q_i \), with \( j \in \{1, 2, \ldots, i\} \) are computed until the number of overlaps is less than one. The value obtained for \( i \) is the number of required digestions.

The results of the statistics approach are scored by an error probability \( e \). The error probability is computed in a combinatorics approach by the number of overlaps \( o \), the number of fragments \( f \) of every digestion.

\[
e(i) := \prod_{j=1}^{i-1} \left( \frac{a_j - o_j}{a_j} \right)^{f_j + 1}
\]

2.4 Theoretical model

Besides this explicit statistics, a theoretical statistics approach that estimates the number of required digestions by use of the genome size \( g \) and the
After the loop, the InFiRe web interface is structured into three steps: selection of genome and replicons; (ii) selection of transposon; (iii) digestion parameters (Fig. 4). These steps reflect the workflow of an InFiRe analysis and should be used in the given order. In the first step, selection of the genome and replicon, all organisms from sequenced bacterial genomes available at the NCBI can be chosen (Fig. 4A). One genome can consist of several replicons, e.g. chromosomes or plasmids. In the second step, the user selects the transposon (Fig. 4B). The software provides three possibilities to enter a transposon: selection from the list of transposons, entering a transposon sequence into the text field or uploading of a FASTA files. Since it is necessary for the InFiRe analysis that the applied enzymes do not cut inside the transposon sequence, the list of permitted restriction enzymes (Enzymes) is assigned after transposon selection. In the third step, the restriction enzymes applied for the digestions can be chosen (Fig. 4C). It should be taken into consideration that one digestion can be performed by more than one enzyme.

3 RESULTS

3.1 InFiRe analysis workflow

The InFiRe web interface is structured into three steps: (i) selection of genome and replicons; (ii) selection of transposon;
The exact size of the fragments was determined by Southern blot hybridization. We successfully applied the algorithm in a case study using the obtained digestion pattern for one of the analyzed L. pneumophila Corby::Tn10 mutants. The approximated size of the hybridized fragments is entered in the corresponding field and analyzed. For determination of the DNA fragment sizes, Southern blot hybridization needs to be performed (Supplementary Materials). The approximated size of the hybridized fragments is entered in the corresponding genome fragment and further genomic information about the sequence, genes and database links are provided. Every location is linked to the NCBI map viewer that provides a visualization of the corresponding genome fragment and further genomic information about the sequence, genes and database links (Fig. 3). Verification of the transposon insertion can be performed by PCR using primer binding inside the transposon sequence and primer binding outside the predicted DNA fragment or two primers binding outside the predicted DNA fragment.

3.2 Application of the InFiRe software and experimental verification of the prediction

We successfully applied the algorithm in a case study using a L. pneumophila Corby mini-Tn10 transposon library. After the screening for L. pneumophila mutants, which are attenuated in intracellular survival within host cells, we received a set of bacterial strains with unknown insertion sites (Shevchuk and Steinert, 2009). We applied the described protocol for the identification of insertion sites in the generated mutants. During the design of the experiment, we chose enzymes that do not cut inside the mini-Tn10 transposon sequence. The number of restriction digestions was statistically calculated and resulted in five digestions. For every digestion, we chose enzymes that do not cut inside the mini-Tn10 transposon sequence and primer binding outside the predicted DNA fragment or two primers binding outside the predicted DNA fragment.

For the determination of insertion sites by standard restriction digestion combined with Southern blot hybridization, the method is applicable to all sequenced organisms and therefore has a great potential to bring benefits in a wide range of applications. In comparison to existing methods, InFiRe has several advantages. First, the method does not require intensive experimental optimization, therefore numerous insertion mutants can be analyzed simultaneously. Second, InFiRe overcomes the difficulties associated with the amplification and sequencing of GC-rich genomic fragments. Lastly, the method allows for the determination of exact position of a transposon insertion in any genome. This is especially important for identification of the transposons, which are integrated in repetitive sequences. Although long repeated sequences are carefully annotated in sequenced genomes, less attention was paid to investigate the biological role of these structures. Thus, InFiRe may also open new ways to analyze the functions of repeated sequences in prokaryotic genomes.

ACKNOWLEDGEMENTS

We thank Stefan Leupold for his help during development of the software and Bernd Hoppe for financial management. We gratefully acknowledge the support of Barbara Schulz in critical proofreading of the manuscript. We are grateful to Stanislav Rosenblit for assistance in web design.

Funding: Deutsche Forschungsgemeinschaft (DFG, grant SFBTR51); Bundesministerium für Bildung und Forschung (BMBF) medical infection genomics (grant no. 0315831A, 0315833A).

Conflict of Interest: none declared.

REFERENCES


Corby chromosome between 762 524 bp and 765 083 bp. The PCR amplification and sequencing confirmed the insertion at position 763 194 bp (Fig. 5).
O. Shevchuk et al.


